DOCUMENT RESUME

ED 380 493 TM 022 854

AUTHOR Pennock-Roman, Maria
TITLE College Major and Gender Differences in the Prediction of College Grades. College Board Report No. 94-2.
INSTITUTION College Board, New York, NY.; Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-94-24
PUB DATE 94
NOTE 25p.
AVAILABLE FROM College Board Publications, Box 886, New York, NY 10101-0886 ($15).
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Achievement; Asian Americans; Black Students; *College Freshmen; Females; *Grade Prediction; Grades (Scholastic); *Grading; Higher Education; High Schools; High School Students; *Majors (Students); Males; Research Methodology; *Sex Differences; Standards; Student Evaluation; White Students
IDENTIFIERS African Americans; Scholastic Aptitude Test

ABSTRACT

Recent studies have found substantial reductions in gender differences in the prediction of academic achievement in colleges when variations in grading standards among courses were taken into account. This project examined gender differences in the prediction of freshman grades after controlling for differential course grading based on college majors. This method involved deriving variables that measured grading leniency using residual scores from the within-gender regressions of freshman grades on high school grades and scores on the Scholastic Aptitude Test for the non-Latino white group. The procedure worked quite well and generalized to other groups not involved in the derivation of the grading leniency scale. Nevertheless, there were modest, sometimes statistically significant, gender differences in prediction that remained after this control variable was introduced into the regressions. The largest and smallest differences for females between actual grades and grades predicted from the males' regressions tended to be found in African American and Asian American groups respectively. The results imply that the use of information on college majors is a reasonable, practical procedure for controlling grading leniency. Thirteen tables present analysis results. (Contains 32 references.) (Author/SLD)

College Board Report No. 94-2

College Major and Gender Differences in the Prediction of College Grades

MARÍA PENNOCK-ROMÁN

The College Board
Educational Excellence for *All* Students

College Board Report No. 94-2
ETS RR No. 94-24

College Major and Gender Differences in the Prediction of College Grades

MARÍA PENNOCK-ROMÁN

College Entrance Examination Board, New York, 1994

Acknowledgments

This research was funded jointly by the College Board and Educational Testing Service. Office space during the completion of this project was provided by Princeton University. Statistical analyses were made possible by Douglas Mills, who aided the author in mastering SAS procedures. The original data sets were maintained with the help of Judith Pollack, Michael Narcowich, and Laura McCamley. Thorough reviews of the manuscript by Lawrence J. Stricker, Nancy Allen, and Ruth B. Ekstrom were very useful and are gratefully acknowledged.

María Pennock-Román is a research scientist in the Division of Applied Measurement Research at ETS.

Additional copies of this report may be obtained from College Board Publications, Box 886, New York, New York 10101-0886. The price is $15.

Printed in the United States of America.

Contents

Abstract

Recent studies have found substantial reductions in gender differences in the prediction of academic achievement in college when variations in grading standards among courses were taken into account. The purpose of this project was to examine gender differences in the prediction of freshman grades after controlling for differential course grading based on college majors. This method involved deriving a variable that measured grading leniency using residual scores from the within-gender regressions of freshman grades on high school grades and scores on the SAT for the non-Latino white group. The procedure worked quite well and generalized to other groups not involved in the derivation of the grading-leniency scale. Nevertheless, there were modest, sometimes statistically significant, gen' differences in prediction that remained after this control variable was introduced into the regressions. The largest and smallest differences for females between actual grades and grades predicted from the males' regressions tended to be found in the African American and Asian American groups, respectively. The results imply that the use of information on college majors is a reasonable, practical procedure for controlling for grading leniency.

Introduction

The objective of this study was to explore sources of possible gender differences in the prediction of college grades at four universities. The analyses focused on separate contributions to these differences by individual predictors: high school grades and SAT scores, both verbal (SAT-V) and mathematical (SAT-M). Of special interest was the extent to which gender differences in predicted versus actual grades persisted after controlling for differential grading standards in the college courses taken by students majoring in different fields of study. All analyses were done separately by racial/ethnic groups within each university to examine variations in the size of gender differences across groups varying in cultural and language background.

Background

A large body of research on a variety of admission tests has shown that the prediction of grades in high school and higher education differs for males and females. Typically, males achieve lower grades than females in high school, college, and law school despite having higher test scores (see reviews by Clark and Grandy 1984, College Board 1988, Linn 1982, and Wilder and Powell 1989; and large, more recent studies by Ramist, Lewis, and McCamley 1994, and Sawyer 1986). In addition, the degree of relationship between predicted and actual grades and the correlations between academic performance in higher education and admission test scores are often stronger for females than for males (Linn 1982; Morgan 1990; Ramist et al. 1994; Sawyer 1986).

Several explanations have been proposed for these findings, including (1) disproportionate enrollment of males in college courses with harsher grading standards, such as the physical sciences; (2) lower percentages of females than of males taking high school science and mathematics courses, thus raising high school grades for females; (3) superior study habits and self-discipline among females; (4) superior writing skills among females; and (5) bias in the tests. Since the major focus here is on grading standards, a full discussion of the other issues (three, four, and five) is beyond the scope of this paper. It can be said briefly that there is evidence partly supporting each of the explanations (Breland and Griswold 1982; Bridgeman 1989; Bridgeman and Wendler 1989, 1991; Bridgeman and Lewis 1991; College Board 1988; Ekstrom, Goertz, and Rock 1988; Elliott and Strenta 1988; Mazzeo, Schmitt, and Bleistein 1989; McCormack and McLeod 1988; Mullis and Jenkins 1988; Ramist et al. 1994; Stricker, Rock, and Burton 1991; Wilder and Powell 1989; Young 1991).

Many studies have established that grading standards vary by field of study and that females tend to gravitate toward the more leniently graded courses and college majors. Typically, more males than females are interested in majoring in engineering and the physical sciences (Grandy 1987a, 1987b), whereas females are more often interested in the humanities and certain social sciences. The distribution of grades in engineering and the physical sciences tends toward higher frequencies in the range of C or below more frequently than the distribution of grades in the humanities and certain social sciences. These differences are found even when previous academic achievement and test scores are taken into account (Elliott and Strenta 1988; Goldman and Hewitt 1975; Goldman et al. 1974; Goldman and Widawski 1976; Strenta and Elliott 1987; Willingham 1985).

A number of studies have found that gender differences in the prediction of college grades (usually in the direction of underprediction of females' grades) are reduced, eliminated, or occasionally reversed when grading leniency was controlled, either by predicting individual course grades (McCormack and McLeod 1988;

Ramist et al. 1994), or by adjusting the cumulative grade-point average (Elliott and Strenta 1988; Stricker et al. 1991; Young 1991). Nevertheless, some gender differences in the prediction of college grades remained statistically significant even after controlling for grading standards (Stricker et al. 1991), but the differences tended to be smaller or nonexistent at more selective colleges with students who had high average composites of high school grades and SAT scores (Ramist et al. 1994). Some authors have argued that test bias (among other explanations) cannot be ruled out (e.g., Elliott and Strenta 1988) because gender differences have persisted in the prediction of individual course grades in psychology (Elliott and Strenta 1988), mathematics (Bridgeman and Wendler 1989, 1991) and a variety of other subject areas (Ramist et al. 1994).

Either method, the prediction of individual course grades or the adjustment of the cumulative grade-point average, used in the aforementioned studies is labor intensive and impractical for routine application in many settings. Both methods depend on the analysis of individual course grades at the undergraduate level, which is fraught with practical difficulties. Not every student takes every course, so for the majority of courses, samples of students enrolled in a particular course are unrepresentative and small. These factors introduce statistical complexities in the analysis of transcript data that are not easily handled by routine procedures. Furthermore, transcript data are not always available in a form readily usable for computer analyses.

As a more practical option, some researchers have categorized grade-point average by schools within a university (Gamache and Novick 1985) or used college major data to control for leniency in course grading in the study of gender-differentiated prediction (Pennock-Román 1990). Unlike transcript data, information on within-university subdivisions or college majors is more accessible. Frequently, university records contain students' intended college majors, or alternatively, the majority of students taking the SAT indicate their intended field of study on the Student Descriptive Questionnaire (SDQ). While the curriculum is less specialized in the freshman year than in the later years of college, differentiation may occur even in the first year. For example, the introductory physics course taken by physics majors may be faster paced and more mathematical than the introductory physics course taken by nonscience majors.

Although little is known about the effectiveness of controlling for grading standards by analyzing grades in terms of subdivisions within institutions, the approach using college majors has shown promising results. Gamache and Novick (1985) did not analyze gender effects on overall grades pooling across college subdivisions; therefore, we cannot tell whether the gender differences they found within each subdivision would have been the same or smaller than in analyses using overall grades pooling all subdivisions. However, using dummy variables to categorize college majors improved the prediction of college grades in studies by Goldman and Hewitt (1976) and Pennock-Román (1990). In these studies, the effects of ethnicity or gender in prediction were reduced but not completely eliminated. Pennock-Román (1990) demonstrated no statistically significant gender effects on freshman college grades at five of six large and prominent universities after controlling for college major in Latino American and non-Latino white groups. It is possible that gender differences might have been completely eliminated if a finer grouping of majors had been achieved and if the classification of majors had been specifically tailored to each institution.[1]

Rationale

In the present investigation, analyses of Pennock-Román's (1990) data were extended in several ways. First, differences in regressions were examined when each of the predictors (high school grades, SAT-V, and SAT-M) were considered jointly versus singly. These analyses evaluated how much each predictor contributed to gender differences in the prediction of grades. For example, the question of possible grade inflation in high school grades for females was addressed by evaluating whether freshman grades were lower than expected given the high school record. Second, slope coefficient differences were examined, whereas the previous study considered only possible intercept differences. Third, the analyses included Asian American and African American samples for whom data were collected and merged but not analyzed in the previous study, which had focused on Latino American and non-Latino white students. Findings on gender differences within Asian American and African American groups are less often available than for the non-Latino white group. In particular, it was expected that the Asian American group would show a more gender-balanced choice of majors in the physical sciences and that gender differences in prediction before ad-

[1]In that study, no further evaluation of the influence of college major on gender differences in prediction was done because the main focus was on the effects of language background on the prediction of college grades. Gender was only one of several control variables.

justing for college majors might be smaller than in other groups.

Finally, the categorization of major fields by institution was improved by grouping majors according to empirically derived measures of grading leniency rather than by similarity of subject matter. In the prior analyses, only a rather crude, four-category classification of college majors was used. The categories were: (1) physical sciences and engineering, (2) biological and health sciences, (3) humanities, prelaw, and social sciences, and (4) business, education, communication, and home economics. There was considerable heterogeneity within these categories (e.g., premedicine, biology, and nursing were grouped together). It is possible that this classification was not a good control for grading leniency at the public university in Texas, the only institution at which a significant gender effect was found. The categorization derived here was expected to control more effectively for grading leniency, thus reducing the gender effect, if this effect was truly an artifact. Gender effects on the prediction of freshman grades were examined before and after taking into account the reclassification of college majors.

Method

Data Source

Four institutions from the Pennock-Román (1990) data set were included: a public university in Texas, a private university in Massachusetts (two freshman classes), and two universities in California, one public and the other private. The original set of institutions included two additional universities that will not be considered here because the relationship between preadmission measures and college grades was atypically low at those institutions. The sample sizes here are smaller than in the previous study for two reasons. One, students lacking any of the predictors—high school grade-point average (HSGPA), SAT-V, or SAT-M—were excluded from the analyses. Second, students reporting that English was not their best language were also excluded from the analyses. Ramist et al. (1994) found that the college grades of non-native speakers of English tended to be underpredicted by test scores. It was desirable to focus here on gender differences among students for whom English is their best language, thus avoiding the additional variability introduced by language background. Information about college majors was directly available from institutional records only for the universities in Texas and Massachusetts; therefore, for the two California universities, responses to the Student Descriptive Questionnaire were used to classify students by major.

Procedure for Categorization of College Majors

In order to identify empirically which college majors had average, substantially easier, or substantially harsher grading standards at each of the four institutions, the first step involved multiple regression analyses for predicting freshman college grade-point average (FGPA) from SAT scores and high school grades. Analyses were run separately at each university for males and females who were non-Latino white. The non-Latino white groups were chosen to classify fields of study because they were the largest groups at each university and they had the greatest variety of majors. The analyses separated groups by gender and race/ethnicity in order to distinguish the effects of college major on FGPA from demographic-group effects. In a regression combining both sexes, it would be difficult to interpret residuals for majors where there was a disproportionate representation of males or females. For example, if physics majors have lower grades, it could be argued that it is not grading standards per se that are tougher for physics majors. Instead, the effect could be due to the disproportionate presence of males with lower FGPAs in comparison with other majors that have more females with higher FGPAs.

The second step was to calculate the residual differences between students' predicted FGPA and their actual FGPA, which were then divided by their standard errors (separate analyses by sex). Then, mean values of the standardized residuals were calculated for groups of students with the same college major, ignoring gender. The assumption here was that the average residual for each major at a given institution is a function of the leniency of grading standards for that major at that institution. This assumption is tenable only if there are a sufficient number of individuals within a category so that other personal idiosyncracies in characteristics that influence freshman grades (e.g., study habits) or statistical "errors" will cancel each other out.

However, there were so many individual categories of majors in the SDQ and the institutional data that some categories included only one student. In order to derive a more stable estimate of grading leniency for fields of study, it was necessary to group related categories of majors with similar residual values. Fields of study were then organized into larger categories similar to the groupings used by the National Research Council (1987, p. 82) in the Annual Survey of Earned Doctorates. The mean standardized residuals were calculated for these broader categories and compared with the means obtained from the finer categorization of majors. The classification of fields of study was refined as an iterative process until each final grouping had at least 6 cases (but typically more than 20) and the residual values for students and subcategories within the grouping were consistent with each other. For example, the broad health sciences grouping was eventually subdivided into three clusters for the two California institutions. There were two large clusters, premedicine and unspecified health sciences, which were kept separate because they had quite different residuals. All other health categories had very few cases. Categories such as preveterinary and predentistry had residuals consistent with premedicine and were assigned to the same category as premedicine. Others with higher residuals than either of the two main clusters were placed into a third health cluster.

The third step was to create a variable (MAJSCAL) that reflected the degree of grading toughness of the student's category of college major at his or her institution as measured by the size of the mean residual for that major. If the mean residual fell in the interval –.0499 to +.0499, it was assigned a value of zero. A mean residual between 0.0500 and 0.1499 was assigned a +1. Fields with mean residuals between –0.1500 and –0.2499 were assigned a –2, and so forth.

The number of categories of majors and the range of MAJSCAL varied by institution. When the sample sizes were large, it was possible to include more categories of majors. For the university in Texas, there were 49 categories of majors, having frequencies from 11 (biochemistry) to 1,160 (prebusiness) in the non-Latino white group. MAJSCAL at this institution ranged from –5 (undetermined, pharmacy, and computer science majors) to +10 (accounting, advertising, marketing, and finance majors). For the university in Massachusetts, there were 32 categories of majors with frequencies ranging from 18 (sociology and criminology) to 916 (unspecified liberal arts) in the non-Latino white group. MAJSCAL at this institution ranged from –7 (engineering majors) to +5 (acting and voice performance majors). For the public university in California, there were 20 categories of majors, ranging in frequency from 13 (history, philosophy, and religion) to 207 (engineering) in the non-Latino white group. MAJSCAL at this institution ranged from –5 (engineering) to +5 (English and education). For the private institution in California, there were 20 categories of majors, ranging in frequency from 6 (nursing and similar health sciences) to 172 (engineering). MAJSCAL at this university ranged from –3 (physical sciences other than engineering) to +6 (foreign languages, history, culture, and religion).

Transformation of Units for Independent and Dependent Variables

In order to preserve significant digits for the raw regression weights in the computer printout, FGPA and HSGPA were multiplied by 10 and SAT scores were divided by 10. Of course, the mean and standard deviation of the transformed FGPA and HSGPA were 10 times larger than the usual values, whereas the mean and standard deviation of the transformed SAT scores were 10 times smaller than the original scores. Correlations and *R*-square values were unaffected by these transformations, but the root mean square error was in the same units as the transformed FGPA, that is, 10 times larger than usual. As intended, the raw regression weights of the transformed SAT scores were 100 times larger when compared with analyses in other studies using untransformed scores. Thus, no significant digits in the regression weights were lost by rounding in the computer printouts, which increased accuracy in the calculation of predicted grades using the male groups' equations.

Regression Analyses

For each gender-by-racial/ethnic group that had at least 40 cases within a university, freshman grades were predicted from high school grades and SAT scores (the standard model). Furthermore, another model was run, adding MAJSCAL (the new variable or major scale) to the standard model. The results of the two models, with and without MAJSCAL, were compared for each group.

To test for gender differences within groups, regression models were run in each racial/ethnic group, pooling males and females. Dummy variables identifying gender were used to test the lines for parallelism (i.e., no interactions or slope coefficient differences) and coincidence (i.e., equal slopes and equal intercepts). All 24 regression models are shown in Table 1. As can be

TABLE 1

Predictors in Regression Models Used to Test for Gender Differences

Model/Control for Grading Leniency	Versions of Model (1) No Gender Terms	(2) Intercept Differences	(3) Intercept and Slope Differences
HSGPA only			
No control	HSGPA	HSGPA + Gender	HSGPA + Gender + G × HSGPA
With control	HSGPA + MAJSCAL	HSGPA + MAJSCAL + Gender	HSGPA + MAJSCAL + Gender + G × HSGPA
SAT-V only			
No control	SAT-V	SAT-V + Gender	SAT-V + Gender + G × SAT-V
With control	SAT-V + MAJSCAL	SAT-V + MAJSCAL + Gender	SAT-V + MAJSCAL + Gender + G × SAT-V
SAT-M only			
No control	SAT-M	SAT-M + Gender	SAT-M + Gender + G × SAT-M
With control	SAT-M + MAJSCAL	SAT-M + MAJSCAL + Gender	SAT-M + MAJSCAL + Gender + G × SAT-M
Standard			
No control	HSGPA + SAT-V + SAT-M	HSGPA + SAT-V + SAT-M + Gender	HSGPA + SAT-V + SAT-M + Gender + G × HSGPA + G × SAT-V + G × SAT-M
With control	HSGPA + SAT-V + SAT-M + MAJSCAL	HSGPA + SAT-V + SAT-M + MAJSCAL + Gender	HSGPA +SAT-V + SAT-M + MAJSCAL + Gender + G × HSGPA + G × SAT-V + G × SAT-M

Note: The interaction terms abbreviate gender as "G" (e.g., G × HSGPA is the interaction between gender and HSGPA). To test whether the regression lines of males and females were coincident, the *R*-squares for version 3 were compared with those of version 1 in the sample that pooled males and females. To test for parallelism, versions 3 and 2 were compared in the pooled sample.

seen, there were eight sets of models, each with three variations. Within a set, the nondummy variables were all the same but the first version had no dummy variables, the second had just the dummy intercept term, and the third had both intercept and slope difference terms.

Three models had only one preadmission measure entered at one time (HSGPA, SAT-V, or SAT-M). The fourth was the standard model that included all three preadmission measures. A second group of four models included all of the same predictors as the first group of four sets, but, in addition, these models included the variable MAJSCAL, which controlled for grading leniency by major. Regression differences were evaluated primarily on the basis of effect sizes; that is, group-difference terms with uniqueness contributions of .01 or larger to the accountable variance (Cohen 1988) were considered nontrivial.

Finally, the eight regression models (without dummy terms) were analyzed with just the males of each racial/ethnic group, and the estimated parameter values for the male groups were used to predict FGPA for the corresponding female group. Average differences between predicted and actual values were calculated for the standard, HSGPA-only, SAT-V-only, and SAT-M-only models, with and without MAJSCAL.

Results

Means, Standard Deviations, and Distribution of Majors

Group sizes, means, and standard deviations are shown for the four institutions by race/ethnicity and gender in Tables 2A, 2B, 2C, and 2D. All groups are shown here, but the small ones, those with fewer than 40 males and 40 females, were not analyzed in the later regressions. Consistent with previous findings, females tended to score lower than males on the SAT-M, although their freshman grades tended to be slightly higher. The Asian American group was the only one for whom the mean FGPA was slightly higher for males at three of the four universities. The variable that reflects grading leniency by major (MAJSCAL) was slightly higher (more positive) for females in the majority of groups.

TABLE 2A

Means and Standard Deviations by Race/Ethnicity and Gender: Texas Institution

Race/Ethnicity Variables	*Mean* Males	*Mean* Females	SD Males	SD Females
NON-LATINO WHITE	N = 2,144	2,004		
FGPA	25.75	26.87	8.80	7.87
HSGPA	34.41	35.62	4.71	3.95
SAT-V	52.73	50.20	8.97	9.21
SAT-M	59.50	53.60	8.83	9.20
MAJSCAL	−0.32	0.21	2.50	2.43
ASIAN AMERICAN	N = 131	106		
FGPA	29.79	28.75	7.83	8.50
HSGPA	36.29	36.57	3.61	3.84
SAT-V	48.31	49.11	11.76	11.69
SAT-M	60.50	56.97	9.90	9.90
MAJSCAL	−0.80	−0.42	2.51	2.54
AFRICAN AMERICAN	N = 92	172		
FGPA	21.22	22.23	8.66	6.67
HSGPA	32.57	34.11	5.64	4.44
SAT-V	45.84	44.10	9.90	8.97
SAT-M	49.63	46.82	10.84	9.52
MAJSCAL	−0.30	−0.16	2.48	2.02
LATINO AMERICAN	N = 290	287		
FGPA	24.67	25.25	8.14	7.76
HSGPA	34.99	35.19	4.53	4.20
SAT-V	47.40	44.11	9.85	10.25
SAT-M	54.61	48.44	9.98	9.63
MAJSCAL	−0.69	−0.37	2.26	2.35
OTHER	N = 14	8		
FGPA	27.61	25.28	8.24	7.01
HSGPA	34.27	37.08	5.12	2.17
SAT-V	51.36	49.25	8.30	7.32
SAT-M	57.93	52.88	8.41	8.84
MAJSCAL	−0.79	0.38	2.29	1.30

Note: Freshman grades (FGPA) and high school grades (HSGPA) were on a scale 0 to 40. SAT scores were on a scale 20 to 80. See Method. The small Other group was not included in regression analyses.

TABLE 2B

Means and Standard Deviations by Race/Ethnicity and Gender: Massachusetts Institution

Race/Ethnicity Variables	*Mean* Males	*Mean* Females	SD Males	SD Females
NON-LATINO WHITE	N = 2,130	2,298		
FGPA	25.87	26.62	6.94	6.29
HSGPA	29.91	30.82	6.16	5.63
SAT-V	52.35	50.70	8.69	8.66
SAT-M	57.80	53.08	9.35	9.18
MAJSCAL	−0.65	0.65	3.44	2.64
ASIAN AMERICAN	N = 154	147		
FGPA	27.81	27.04	7.63	6.68
HSGPA	33.77	33.67	5.50	6.38
SAT-V	54.72	51.68	8.86	9.86
SAT-M	63.23	57.31	7.74	9.51
MAJSCAL	−1.72	0.05	3.89	3.11
AFRICAN AMERICAN	N = 66	112		
FGPA	21.97	23.07	6.01	5.99
HSGPA	27.92	29.09	5.68	5.66
SAT-V	47.17	45.36	9.41	8.51
SAT-M	49.79	45.92	10.08	8.75
MAJSCAL	−1.31	0.05	3.23	2.33
LATINO AMERICAN	N = 62	54		
FGPA	23.01	24.13	5.73	6.75
HSGPA	31.18	32.27	5.53	4.37
SAT-V	49.05	48.17	9.10	8.66
SAT-M	53.02	49.78	8.06	8.37
MAJSCAL	−1.40	0.15	3.38	2.81
OTHER	N = 53	51		
FGPA	24.64	25.83	8.00	7.03
HSGPA	30.54	31.05	6.74	5.24
SAT-V	50.32	46.88	9.09	10.49
SAT-M	58.53	51.59	9.19	10.43
MAJSCAL	−1.08	0.25	3.44	2.49
MISSING RACE/ETHNICITY	N = 508	495		
FGPA	23.90	25.31	7.31	6.57
HSGPA	26.76	27.63	6.20	5.86
SAT-V	48.68	48.29	9.59	8.60
SAT-M	55.23	50.73	9.98	9.04
MAJSCAL	−0.22	0.55	2.77	2.06

Note: Freshman grades (FGPA) and high school grades (HSGPA) were on a scale 0 to 40. SAT scores were on a scale 20 to 80. See Method.

These mean differences are consistent with patterns of choice of major field. When quantitative majors were grouped into one category (including mathematics, engineering, computer science, physical sciences, earth sciences, and ocean sciences), the percentage of males majoring in this category was 14 to 31 percentage points higher than the percentage of females of the same race/ethnicity majoring in that category. There were two exceptions, however. At the private California institution, the percentages of males versus females majoring in the sciences were nearly equal in the Latino American group (42 percent versus 40 percent, respectively) and in the Asian American group (29 percent

TABLE 2C

Means and Standard Deviations by Race/Ethnicity and Gender: California Public Institution

Race/Ethnicity	*Mean*		*SD*	
Variables	*Males*	*Females*	*Males*	*Females*
NON-LATINO WHITE	N = 735	537		
FGPA	29.67	30.19	5.48	4.67
HSGPA	36.62	36.99	3.49	2.80
SAT-V	56.50	55.26	8.90	8.68
SAT-M	64.55	57.83	8.20	7.88
MAJSCAL	-0.51	0.86	2.76	2.27
ASIAN AMERICAN	N = 278	283		
FGPA	29.99	29.36	5.47	5.34
HSGPA	37.07	37.21	3.12	3.26
SAT-V	52.01	51.54	9.67	9.41
SAT-M	63.87	58.14	8.96	9.15
MAJSCAL	-1.18	-0.03	2.75	2.32
AFRICAN AMERICAN	N = 23	37		
FGPA	22.03	22.60	7.31	6.47
HSGPA	33.26	33.43	5.04	3.89
SAT-V	49.74	42.95	9.84	9.53
SAT-M	54.52	44.97	9.54	9.26
MAJSCAL	-0.74	-0.16	2.68	2.65
LATINO AMERICAN	N = 42	25		
FGPA	26.28	27.02	5.17	4.57
HSGPA	35.20	35.82	3.81	3.24
SAT-V	48.05	49.56	10.24	9.50
SAT-M	54.74	48.12	9.55	8.07
MAJSCAL	0.00	0.92	2.71	2.66
OTHER/MISSING	N = 56	54		
FGPA	29.34	29.26	6.46	4.71
HSGPA	36.65	36.61	3.55	2.75
SAT-V	55.63	51.28	9.81	8.79
SAT-M	61.46	54.11	9.65	9.01
MAJSCAL	-0.70	0.59	2.88	1.98

Note: Freshman grades (FGPA) and high school grades (HSGPA) were on a scale 0 to 40. SAT scores were on a scale 20 to 80. See Method. The small African American and Latino American groups were not included in the regression analyses.

TABLE 2D

Means and Standard Deviations by Race/Ethnicity and Gender: California Private Institution

Race/Ethnicity	*Mean*		*SD*	
Variables	*Males*	*Females*	*Males*	*Females*
NON-LATINO WHITE	N = 510	380		
FGPA	32.86	32.97	3.96	3.60
HSGPA	38.14	38.28	2.43	2.33
SAT-V	61.27	60.68	7.61	7.20
SAT-M	68.16	63.31	7.11	7.74
MAJSCAL	-0.42	0.52	2.09	2.33
ASIAN AMERICAN	N = 63	44		
FGPA	33.35	33.43	3.46	3.07
HSGPA	38.72	39.01	1.65	1.61
SAT-V	63.71	62.68	6.42	7.21
SAT-M	71.00	67.23	6.58	5.71
MAJSCAL	-0.73	-0.34	1.58	1.80
AFRICAN AMERICAN	N = 59	57		
FGPA	27.97	29.37	3.89	3.75
HSGPA	35.05	35.53	4.20	3.58
SAT-V	54.29	55.00	8.10	7.73
SAT-M	58.98	56.93	9.50	7.63
MAJSCAL	-0.81	-0.47	1.73	1.74
LATINO AMERICAN	N = 64	42		
FGPA	28.45	28.90	3.76	3.81
HSGPA	37.47	37.81	2.79	1.94
SAT-V	53.36	52.74	9.13	7.48
SAT-M	60.56	56.14	8.18	7.12
MAJSCAL	-0.81	-0.48	1.58	1.84
OTHER/MISSING	N = 115	86		
FGPA	32.26	32.97	3.56	3.54
HSGPA	36.30	36.30	3.84	3.76
SAT-V	60.08	59.93	8.94	8.18
SAT-M	66.93	62.15	7.39	8.22
MAJSCAL	0.07	0.24	1.29	1.53

Note: Freshman grades (FGPA) and high school grades (HSGPA) were on a scale 0 to 40. SAT scores were on a scale 20 to 80. See Method.

versus 25 percent, respectively). In contrast, the largest gender differences in this category of majors were found among Asian Americans (47 percent versus 16 percent) and Latino Americans (37 percent versus 9 percent) at the university in Texas, and the non-Latino white students at the public California institution (41 percent versus 16 percent). Hence, contrary to the author's expectation, the pattern of major choices in the Asian American group did not show greater gender balance among quantitative majors compared with other ethnic groups.

Regression Analyses with and without MAJSCAL

Results for two sets of regressions of FGPA on HSGPA and SAT scores using the standard model are shown in

TABLE 3A

Standard Model Prediction of Freshman Grades by Gender and Race/Ethnicity: Texas Institution

Race/Ethnicity Gender	Without MAJSCAL N	Without MAJSCAL Root MSE	Without MAJSCAL R-Square	With MAJSCAL Root MSE	With MAJSCAL R-Square	With MAJSCAL Contribution to R-Square	With MAJSCAL P-Value
NON-LATINO WHITE							
Males	2,144	7.7320	.2298	7.4412	.2870	.0572	.0001
Females	2,004	6.7354	.2682	6.5652	.3051	.0369	.0001
ASIAN AMERICAN							
Males	131	5.9216	.4420	5.9136	.44	.0059	ns*
Females	106	6.0155	.5138	5.9031	.5364	.0226	0288
AFRICAN AMERICAN							
Males	92	8.2810	.1151	7.8422	.2154	.1003	.0013
Females	172	6.0713	.1857	6.0402	.1989	.0131	.1001
LATINO AMERICAN							
Males	290	7.4918	.1616	7.4696	.1695	.0079	.1013
Females	287	6.7222	.2568	6.6828	.2681	.0113	.0380

Note: The sizes of the root mean square errors reflect the change to a scale of 0 to 40 for FGPA (see Method). Differences between the two R squares do not always agree exactly with the contribution to R-square because of rounding errors.
*ns = nonsignificant, indicating p values exceed .1100, a value not considered even marginally significant.

TABLE 3B

Standard Model Prediction of Freshman Grades by Gender and Race/Ethnicity: Massachusetts Institution

Race/Ethnicity Gender	Without MAJSCAL N	Without MAJSCAL Root MSE	Without MAJSCAL R-Square	With MAJSCAL Root MSE	With MAJSCAL R-Square	With MAJSCAL Contribution to R-Square	With MAJSCAL P-Value
NON-LATINO WHITE							
Males	2,130	6.2618	.1875	5.9711	.2615	.0740	.0001
Females	2,298	5.5214	.2300	5.2457	.3053	.0753	.0001
ASIAN AMERICAN							
Males	154	6.5173	.2854	6.2923	.3383	.0529	.0007
Females	157	5.8960	.2368	5.8160	.2626	.0258	.0275
AFRICAN AMERICAN							
Males	66	6.0239	.0416	6.0205	.0582	.0165	ns*
Females	112	5.8029	.0857	5.6060	.1564	.0707	.0034
LATINO AMERICAN							
Males	62	5.4887	.1266	5.1089	.2563	.1297	.0026
Females	54	6.4674	.1335	6.3132	.1909	.0574	.0683
OTHER							
Males	53	6.2494	.4245	6.0690	.4683	.0438	.0524
Females	51	5.3189	.4625	5.1666	.5037	.0411	.0570
MISSING RACE/ETHNICITY							
Males	508	6.6713	.1711	6.4788	.2198	.0487	.0001
Females	495	5.8383	.2162	5.7043	.2533	.0371	.0001

Note: The sizes of the root mean square errors reflect the change to a scale of 0 to 40 for FGPA (see Method). Differences between the two R-squares do not always agree exactly with the contribution to R square because of rounding errors.
*ns = nonsignificant, indicating p values exceed .1100, a value not considered even marginally significant.

13

TABLE 3C

Standard Model Prediction of Freshman Grades by Gender and Race/Ethnicity: California Institutions

Race/Ethnicity Gender	Without MAJSCAL N	Root MSE	R-Square	With MAJSCAL Root MSE	R-Square	Contribution to R-Square	P-Value
CALIFORNIA: PUBLIC							
NON-LATINO WHITE							
Males	735	4.9815	0.1761	4.7631	0.2478	.0717	.0001
Females	537	4.3258	0.1467	4.1793	0.2050	.0583	.0001
ASIAN AMERICAN							
Males	278	4.8522	0.2227	4.8605	0.2229	.0002	ns*
Females	283	4.3581	0.3407	4.2968	0.3614	.0207	.0029
OTHER/MISSING RACE/ETHNICITY							
Males	56	6.2079	0.1280	6.1946	0.1484	.0204	ns*
Females	54	3.9267	0.3438	3.9649	0.3443	.0006	ns*
CALIFORNIA: PRIVATE							
NON-LATINO WHITE							
Males	510	3.5158	0.2181	3.4397	0.2530	.0350	.0001
Females	380	3.3573	0.1363	3.2242	0.2055	.0692	.0001
ASIAN AMERICAN							
Males	63	3.0659	0.2509	3.0198	0.2855	.0347	.0988
Females	44	2.7207	0.2686	2.5856	0.3559	.0873	.0269
AFRICAN AMERICAN							
Males	59	3.7260	0.1283	3.5918	0.2047	.0764	.0267
Females	57	3.1514	0.3314	3.1662	0.3379	.0064	ns*
LATINO AMERICAN							
Males	64	3.5823	0.1347	3.5031	0.1863	.0516	.0579
Females	42	3.5687	0.1848	3.6130	0.1864	.0016	ns*
OTHER/MISSING RACE/ETHNICITY							
Males	115	3.1420	0.2423	3.1559	0.2424	.0002	ns*
Females	86	3.1320	0.2432	3.1004	0.2675	.0242	.1055

Note: The sizes of the root mean square errors reflect the change to a scale of 0 to 40 for FGPA (see Method). Differences between the two *R*-squares do not always agree exactly with the contribution to *R*-square because of rounding errors.
*ns = nonsignificant, indicating *p* values exceed .1100, a value not considered even marginally significant.

Tables 3A, 3B, and 3C for gender by racial/ethnic groups having at least 40 cases. One set of analyses included only HSGPA and SAT scores as predictors, and the second set of regressions included these same three predictors plus the additional variable MAJSCAL that controls for grading leniency by college major. The root mean square errors and *R*-squares are shown for both models; further, the contribution to *R*-square by the addition of the MAJSCAL variable is reported. Note that Table 3C contains results for both institutions in California.

As these tables show, for the majority of groups in all institutions, there was a substantially greater *R*-square for the model that contains MAJSCAL. Such an improvement would have to occur by necessity in the non-Latino white groups because residual values from the regressions in these groups were the basis for deriving the MAJSCAL variable. However, the MAJSCAL derivation did not depend at all on residual values for the Asian American, African American, and Latino American groups; therefore, its application in these groups can be considered a cross validation of its usefulness. The results show that the increases in *R*-square were fairly large in the majority of groups other than the non-Latino white groups. Thus, the method appears to have cross validated about as well as can be expected.

Male and Female Differences in Regressions

Differences in the regressions for the two gender groups are summarized for non-Latino white, Asian American,

TABLE 4

Gender Differences after Controlling for Majors: Contributions to R-Square

	Standard Model		*HSGPA-Only*		*SAT-V Only*		*SAT-M Only*	
Group/School	*Intercept*	*Three Slopes*	*Intercept*	*Slope*	*Intercept*	*Slope*	*Intercept*	*Slope*
NON-LATINO WHITE								
Texas	.0082[a,e]	.0007	.0000	.0000	.0072[a,e]	.0005	.0231[a,e]	.0002
Massachusetts	.0001	.0005	.0012[b,d]	.0000	.0011[c,e]	.0000	.0057[a,e]	.0005
California: Public	.0005	.0038	.0011	.0001	.0003	.0022	.0072[b,e]	.0057[b,d]
California: Private	.0004	.0088[c]	.0011	.0016	.0001	.0009	.0041[b,e]	.0101[b,d]
AFRICAN AMERICAN								
Texas	.0047	.0020	.0004	.0003	.0069	.0000	.0096	.0005
Massachusetts	.0025	.0295	.0010	.0023	.0038	.0089	.0041	.0044
California: Private	.0272[c,e]	.0316	.0211[e]	.0188[e]	.0228[e]	.0010	.0373[c,e]	.0055
LATINO AMERICAN								
Texas	.0047	.0024	.0004	.0023	.0058	.0026	.0105[c,e]	.0022
Massachusetts	.0000	.0410	.0003	.0028	.0020	.0268[e]	.0051	.0350[c,e]
California: Private	.0014	.0100	.0006	.0047	.0038	.0034	.0066	.0004
ASIAN AMERICAN								
Texas	.0020	.0192[c]	.0092	.0012	.0077	.0135[c]	.0002	.0000
Massachusetts	.0004	.0102	.0111[c,d]	.0143[c,d]	.0013	.0004	.0006	.0105[c,d]
California: Public	.0005	.0062	.0042	.0025	.0020	.0047	.0050	.0000
California: Private	.0026	.0638[c]	.0010	.0167[d]	.0004	.0055	.0085	.0177[e]

Note: The direction of differences for models having only one predictor is shown only for statistically significant results or contributions to R square exceeding .01.
[a]Significant at the .0001 level.
[b]Significant at the .01 level.
[c]Significant at the .05 level.
[d]Higher value for males.
[e]Higher value for females.

African American, and Latino American students in Tables 4 and 5A, 5B, and 5C. Table 4 reports statistical tests to evaluate intercept and slope coefficient differences after the control for major was introduced. Tables 5A to 5C compare actual and predicted values for mean FGPA for females, by race/ethnicity. Note that there are two racial/ethnic groups in Table 5B. The original scale for FGPA with a maximum value of 4.00 was used in these tables. The males' equations were used to derive predicted FGPA for females. The first four columns show the results for models not including a control for leniency in grading standards by major. The model referred to as the "Standard 3" used HSGPA and SAT scores as the three predictors, whereas each of the other three regression models shown used only one of these variables as a predictor. The second set of four columns shows the results for parallel analyses, this time based on regressions that included the same corresponding variables plus MAJSCAL. (Of course, values for males are not shown since the use of the males' equations guarantees perfect agreement between the actual and predicted values at the mean.)

Standard Model

The results in Tables 4 and 5A, 5B, and 5C showing differences between males and females in the standard model were very consistent across groups and universities in terms of general pattern. Although gender effects tended to be somewhat smaller when MAJSCAL was added, the pattern of differences for a given group was unaltered by the inclusion of MAJSCAL. To avoid redundancy, only the statistical tests with the control for grading leniency are shown in Table 4. For example, before the inclusion of MAJSCAL, there were only two intercept-difference terms that contributed more than .01 to R-square: the non-Latino white group at the university in Texas and the African American group at the private California institution. These two terms were reduced, respectively, from .0124 to .0082 and from .0315 to .0272; nevertheless, they remained statistically

TABLE 5A

Mean Predicted FGPA for Non-Latino White Female Students Using Male Students' Regressions

	Predictors in Regression Models							
	Without MAJSCAL				*Each Below Plus MAJSCAL*			
University	*Standard 3 Combined*	*HSGPA*	*SAT-V*	*SAT-M*	*Standard 3 Combined*	*HSGPA*	*SAT-V*	*SAT-M*
TEXAS	Actual Mean FGPA = 2.687							
Predicted FGPA	2.502	2.659	2.503	2.382	2.541	2.704	2.555	2.431
Actual minus predicted	0.185	0.028	0.183	0.305	0.146	−0.017	0.132	0.256
MASSACHUSETTS	Actual Mean FGPA = 2.662							
Predicted FGPA	2.572	2.627	2.548	2.484	2.634	2.702	2.605	2.541
Actual minus predicted	0.089	0.035	0.114	0.178	0.027	−0.040	0.056	0.120
CALIFORNIA: PUBLIC	Actual Mean FGPA = 3.019							
Predicted FGPA	2.933	2.984	2.943	2.852	2.975	3.053	2.998	2.897
Actual minus predicted	0.086	0.035	0.076	0.167	0.044	−0.034	0.021	0.122
CALIFORNIA: PRIVATE	Actual Mean FGPA = 3.297							
Predicted FGPA	3.240	3.292	3.274	3.195	3.260	3.315	3.299	3.218
Actual minus predicted	0.057	0.005	0.023	0.102	0.037	−0.018	−0.016	0.079

Note: Values for FGPA, predictions, and differences are in the original units (scale 0 to 4). Mean predicted FGPA and actual mean FGPA are not shown for males because the two are identical except for rounding errors, since the males' equations were used to calculate predicted values.

significant and the amount of underprediction of actual grades was still modestly large for each (0.146 and 0.128, respectively). The other groups showed no large or statistically significant intercept differences. Some slope coefficient dissimilarities were found but there were no consistent patterns across groups and universities in terms of the variable involved or the direction of the gender difference.

Overall, the actual freshman grades of females were higher than the values predicted using the male students' regression equations, both before and after including the control variable for grading leniency (MAJSCAL) in the regressions. There were only three exceptions out of 14 contrasts. Predicted values were higher than actual FGPA for females in the Latino American group at the Massachusetts university and in the Asian American groups at both the Texas and Massachusetts universities. These differences were less than 0.08 grade-point units in absolute value.

HSGPA-Only Model

Findings for this model were similar to the standard model results for African American and Latino American students in that FGPA tended to be higher than predicted when the males' equations were used, both before and after including MAJSCAL. The effects were trivially small and nonsignificant except for the African American group at the California private institution, which also had a large slope coefficient difference (HSGPA more correlated with FGPA for females). With few exceptions, the divergence between actual and predicted grades was smaller with the HSGPA-only model than with the standard model.

The pattern of differences was reversed for Asian American students with this model—the males' equations tended to overpredict the grades of females, and the differences tended to become larger in a negative direction when MAJSCAL was added. Differences ranged from −0.032 to −0.164 grade-point units when MAJSCAL was included in the regression model. These figures tended to be slightly larger in absolute value than for the standard model. Among non-Latino white students, there was underprediction of females' grades before including MAJSCAL, but the direction of the difference reversed when MAJSCAL was added to the equations. The degree of overprediction ranged from −0.017 to −0.040 grade-point units when MAJSCAL was included. These differences tended to be smaller in absolute value than the standard model differences for the non-Latino white group. Regardless of the direction of the effect, gender dissimilarities in intercepts for Asian Americans and non-Latino whites were not statistically significant except at the Massachusetts univer-

TABLE 5B

Mean Predicted FGPA for African American and Latino American Female Students Using Male Students' Regressions

	Predictors in Regression Models							
	Without MAJSCAL				*Each Below Plus MAJSCAL*			
Group *University*	*Standard 3 Combined*	*HSGPA*	*SAT-V*	*SAT-M*	*Standard 3 Combined*	*HSGPA*	*SAT-V*	*SAT-M*
AFRICAN AMERICAN								
TEXAS	Actual Mean FGPA = 2.223							
Predicted FGPA	2.119	2.184	2.081	2.065	2.135	2.203	2.098	2.079
Actual minus predicted	0.104	0.039	0.142	0.158	0.088	0.020	0.125	0.144
MASSACHUSETTS	Actual Mean FGPA = 2.307							
Predicted FGPA	2.190	2.218	2.189	2.166	2.215	2.240	2.204	2.186
Actual minus predicted	0.117	0.089	0.118	0.141	0.092	0.067	0.103	0.121
CALIFORNIA: PRIVATE	Actual Mean FGPA = 2.937							
Predicted FGPA	2.795	2.804	2.807	2.774	2.809	2.826	2.825	2.799
Actual minus predicted	0.142	0.133	0.130	0.163	0.128	0.111	0.112	0.138
LATINO AMERICAN								
TEXAS	Actual Mean FGPA = 2.525							
Predicted FGPA	2.410	2.479	2.405	2.372	2.413	2.491	2.416	2.378
Actual minus predicted	0.115	0.046	0.120	0.153	0.112	0.034	0.109	0.147
MASSACHUSETTS	Actual Mean FGPA = 2.413							
Predicted FGPA	2.361	2.338	2.307	2.313	2.484	2.455	2.357	2.344
Actual minus predicted	0.052	0.075	0.106	0.100	−0.071	−0.042	0.056	0.069
CALIFORNIA: PRIVATE	Actual Mean FGPA = 2.890							
Predicted FGPA	2.858	2.856	2.838	2.821	2.867	2.876	2.857	2.836
Actual minus predicted	0.032	0.034	0.052	0.069	0.023	0.014	0.033	0.054

Note: Values for FGPA, predictions, and differences are in the original units (scale 0 to 4). Mean predicted FGPA and actual mean FGPA are not shown for males because the two are identical except for rounding errors, since the males' equations were used to calculate predicted values.

sity and these effects were small. There was only one intercept difference (at the Massachusetts institution) that contributed more than .01 to R-square for Asian Americans and the only significant effect for the non-Latino white group had a trivially small contribution to R-square (.0012). At two institutions, the Massachusetts and private California universities, HSGPA was more correlated with FGPA for Asian American males and there were fairly large slope coefficient differences.

SAT-V-Only Model

Overall, the results with the SAT-V-only model for the first three groups were quite similar to the findings with the standard model in that females' grades were underpredicted by the males' equations. The only large intercept difference occurred for the African American group at the private California institution. One slope coefficient difference, for the Latino American group at the Massachusetts university, had a contribution to R-square larger than .01 (higher correlation between SAT-V and FGPA for females). The amount of underprediction of females' grades was about the same or slightly larger than with the standard model (ranging from −0.016 to 0.135) when grading leniency was controlled.

The grades of Asian American females tended to be overpredicted by the males' equations when MAJSCAL was included. These differences ranged from −0.125 to 0.031 without MAJSCAL and from −0.136 to 0.021 after including MAJSCAL. None of the intercept differences contributed more than .01 to R-square; there was one fairly large slope coefficient difference at the Texas university, in that SAT-V was more highly correlated with FGPA for females.

TABLE 5c

Mean Predicted FGPA for Asian American Female Students Using Male Students' Regressions

	Predictors in Regression Models							
	Without MAJSCAL				*Each Below Plus MAJSCAL*			
University	*Standard 3 Combined*	*HSGPA*	*SAT-V*	*SAT-M*	*Standard 3 Combined*	*HSGPA*	*SAT-V*	*SAT-M*
TEXAS	Actual Mean FGPA = 2.875							
Predicted FGPA	2.896	3.012	3.000	2.824	2.904	3.027	3.011	2.844
Actual minus predicted	−0.021	−0.137	−0.125	0.051	−0.029	−0.152	−0.136	[illegible]
MASSACHUSETTS	Actual Mean FGPA = 2.704							
Predicted FGPA	2.630	2.775	2.688	2.534	2.728	2.868	2.752	2.611
Actual minus predicted	0.074	−0.071	0.016	0.170	−0.024	−0.164	−0.048	0.093
CALIFORNIA: PUBLIC	Actual Mean FGPA = 2.934							
Predicted FGPA	2.882	3.005	2.990	2.846	2.880	2.992	2.969	2.841
Actual minus predicted	0.052	−0.071	−0.056	0.088	0.054	−0.058	−0.035	0.093
CALIFORNIA: PRIVATE	Actual Mean FGPA = 3.343							
Predicted FGPA	3.324	3.354	3.312	3.287	3.339	3.375	3.322	3.298
Actual minus predicted	0.019	−0.011	0.031	0.056	0.004	−0.032	0.021	0.045

Note: Values for FGPA, predictions, and differences are in the original units (scale 0 to 4). Mean predicted FGPA and actual mean FGPA are not shown for males because the two are identical except for rounding errors, since the males' equations were used to calculate predicted values.

SAT-M-Only Model

This model showed consistent and fairly large underprediction, on average, of females' grades for all groups, including Asian Americans, at all universities. The equations not including MAJSCAL underpredicted females' grades by 0.051 to 0.305 grade-point units, and those including MAJSCAL underpredicted females' grades by 0.045 to 0.256 grade-point units. Intercept differences in the models not including MAJSCAL had contributions to R-square larger than .01 for 9 of 14 contrasts. These differences were smaller in the models including MAJSCAL, but three contrasts had contributions to R-square larger than .01—the non-Latino white group at the Texas university, the African American group at the private California university, and the Latino American group at the Texas university. Some slope coefficient differences were found, but they were not consistent in direction.

Comparison of R-Squares across Models

Naturally, the model that had the most accurate prediction (i.e., the highest R-squares and lowest root mean square errors) in every group at all institutions was the standard model. For analyses where males and females were pooled together, R-squares for this model ranged from .0864 to .4703 for different racial/ethnic groups and universities when MAJSCAL was included in the analysis. In contrast, R-squares for the HSGPA-only, SAT-V-only, and SAT-M-only models ranged from .0558 to .3409, .0373 to .2423, and .0318 to .3066, respectively. The addition of SAT-V and SAT-M jointly to HSGPA in the standard model produced an increment in R-square that ranged from .02 to .17 for different groups and universities.

Discussion

This study addressed three basic issues: How effectively can the use of information on college majors control for differential grading practices across fields of study? What are the relative contributions of each predictor variable (HSGPA, SAT-V, SAT-M) separately and in combination to gender differences in the prediction of college grades? Can differential grading practices across fields of study account for variations in gender differences across racial/ethnic groups and across universities?

Effectiveness of College Majors as a Control for Grading Leniency

Although many prior studies have established that leniency in grading varies by subject area, the question remains of how to control for these effects in a practical yet effective way. Because variations in the leniency of grading practices do not reflect real differences in achievement, these scale differences are an important nuisance factor that needs to be controlled. The most precise, exact methods to control it have involved use of individual course grades from transcripts, an approach that is not feasible in most studies.

The present study investigated to what extent one could control for differential grading across courses by using information on college majors instead of individual course grades. Information on college majors is generally more readily obtainable than full transcript data. Furthermore, analyses involving college majors are relatively straightforward and avoid the complexities of having unequal, and often very small, groups of students in individual courses. The categorization of majors was different from that used in prior studies in that it was carried out in a fashion tailored specifically to each institution.

It was found that introducing the variable MAJSCAL, which controlled for differential grades across majors, did increase predictive accuracy for nearly all groups and did reduce intercept differences and the amount of underprediction of females' grades. The present study confirmed prior findings with regard to differential grading practices. Hence, there is ample evidence from this study and others that college grades differ in scale across fields of study and that variations in grading leniency contribute to variations in grade-point average. Since males and females are unequally distributed among fields that differ in grading leniency, such variations need to be controlled if we are to examine gender differences in the prediction of college grades.

Despite substantial improvement in the accuracy of prediction, this method worked no better than the earlier categorization of majors in terms of reducing gender differences. The intercept–difference term here for the Texas university using the standard model was about the same size and still as statistically significant as in the earlier analysis (Pennock-Román 1990).

The near equality of results for the dummy-variable approach (Pennock-Román 1990) and the use of MAJSCAL in this study to control for variations in grading leniency by major suggests that the most important distinctions can be preserved by using three broad categories: quantitative sciences, biological sciences, and nonquantitative nonscience fields. For example, in analyses of combined Latino American and non-Latino white groups, Pennock-Román (1990, Table 3.15) found that humanities, social sciences, business, and education majors were more leniently graded than physical sciences and engineering majors at all six institutions studied, but the biological/health sciences showed less consistent results. The findings from the present investigation and Pennock-Román (1990) are consistent with many studies that have shown large contrasts in grading leniency between quantitative and nonquantitative majors (Elliott and Strenta 1988; Goldman and Hewitt 1975; Goldman et al. 1974).

Another issue regarding the categorization of majors concerned the use of students' responses to the SDQ at the time that they were taking the SAT versus the institutional records of students' majors. Judging by the size of the contribution to R-square of the variable MAJSCAL, the procedure worked about as well for the two California universities using SDQ data as it did using information provided by the Texas and Massachusetts universities. However, it cannot be said from this study whether the two sources of information are interchangeable or equally valid. In future studies on the efficacy of using controls for college major, these results should be confirmed by comparing controls based on information from the SDQ with controls based on actual institutional records on declared majors at the same university.

Comparisons Across Regression Models

Another goal of this study was to separate as much as possible the degree of differential prediction by gender attributable to the individual predictors: high school grades, SAT-V, and SAT-M. Although the standard, recommended practice for admission committees is to use these variables in combination rather than separately, these analyses can help pinpoint the source(s) of underprediction for female students. For example, it is important to know whether HSGPA underpredicts or overpredicts college grades. Because HSGPA is often higher for females than for males with the same test scores, one could interpret the findings in several ways. One could speculate that tests are biased against females, that females are more diligent students, that teachers' grades may be biased against males, or that females avoid high school science and mathematics courses with tough grading standards. There is evidence that a slightly higher percentage of males than females take high school mathematics and science courses (see the High School and Beyond survey, Ekstrom et al. 1988, and the National Assessment of Educational Progress, Mullis and Jenkins 1988). If patterns of high school course-taking or teachers' biases raise high school grades for females, then their grades would be inflated by factors unrelated to later performance in college. Thus, if this were true, we would expect to find larger intercept differences for the HSGPA-only model as compared with the regressions involving test scores, particularly when controlling for differential grading standards in college courses by college major. The direction of the expected difference, if high school grades were inflated for females, would be that females' actual college grades would be lower than those predicted from the males' equations.

The results do not support the hypothesis of inflated high school grades for females for the majority of groups. For the African American and Latino American groups, the differences between predicted and actual mean FGPA for females after controlling for major were nearly always in the positive direction (underprediction of college grades). There was only one negative value—the Latino American group at the university in Massachusetts (–0.042)—and the intercept and slope coefficients jointly contributed less than .003 in that case. Among the non-Latino white students, the differences were consistently negative but closer to zero than the differences in all other models; all absolute values were less than 0.040 and the intercept and slope coefficient contrasts had no joint contributions to R-square larger than .003 in this group. The pattern of results for the Asian American group was somewhat different, as discussed later.

On the other hand, there was clear evidence that the largest underprediction, on average, of female students' grades resulted with the SAT-M-only model and this underprediction persisted after controlling for grading leniency by major. This underestimation of females' grades using the males' equations occurred even when there was a high proportion of females in quantitative majors that nearly matched the proportion of males in those fields. At the private California institution, the percentages of males versus females majoring in the sciences were nearly equal in the Latino American group (42 percent versus 40 percent, respectively) and the Asian American group (29 percent versus 25 percent, respectively), yet the differences between actual and predicted grades for females after controlling for MAJSCAL were 0.054 for the Latino American group and 0.045 for the Asian American group. In contrast, differences between the actual and predicted grades for females in these groups using the HSGPA-only model were closer to zero (0.014 and –0.032, respectively). These results are consistent with those of Bridgeman and Wendler (1991), who found that SAT-M underpredicted female students' grades in individual mathematics courses. Thus, SAT-M underpredicted the academic achievement of female students in mathematics and in the broad spectrum of courses taken by science and nonscience majors.

For the standard model and the SAT-V-only models, gender differences were typically small after controlling for leniency in grading standards, although sometimes still statistically significant. The small residual underprediction of females' grades was consistent with many studies (Linn 1982; Ramist et al. 1994; Stricker et al. 1991). Other variables not included here, such as study habits and essay writing skills, may account for these differences.

Aside from the issue of under- and overprediction, the models were also evaluated in terms of the degree of relationship between actual and predicted values (R-square). The standard model showed the best overall predictive ability in terms of smaller root mean square errors and multiple R-squares than other models. Thus, in agreement with the bulk of research in this area, the results of this study support the use of the standard model over others because the differences between actual and predicted grades were not much larger than with the HSGPA-only model, and the R-squares were more substantial.

Gender Differences by Race/Ethnicity

Another major goal of this investigation was to extend the analyses in the previous report to African American and Asian American groups at the same institutions because data on gender differences for these groups are less frequently reported. Owing to subgroup differences in language background and other variables, the examination of subgroup variations may provide clues for sources of cross-cultural influences on gender differences. For example, a higher incidence of bilingualism in a group may reduce the female advantage in essay writing. If subgroups vary in degree and direction of gender differences, then ignoring such differences by combining all students in a single group may lead to inconsistent results across universities that differ in racial/ethnic composition. In this study, the combined groups were quite diverse and could not be considered equivalent; whereas the public institution in California had a high percentage of Asian Americans, the university in Texas had a high percentage of Latino Americans. Hence, the analysis by race/ethnicity may be more useful for making generalizations across institutions because it separates population group differences from university-specific characteristics.

In this study it was found that while the pattern of gender differences for the African American group was similar to that of the non-Latino white group, the degree of underprediction of females' grades was slightly larger in the former group (Table 5B). In contrast, the Asian American group showed slightly smaller differences between females' actual grades and predicted grades than those found among the non-Latino white group. Gender differences were also smaller for the Latino American group than for the non-Latino white group, as found in the earlier analyses (Pennock-Román 1990).

Reduced Gender Differences in the Asian American Group

Unlike other groups, actual FGPA for Asian American females was found to be lower than predicted FGPA using the males' parameter estimates for the HSGPA-only equation without MAJSCAL. At three of four universities, these differences became slightly larger in absolute value when major was controlled, but only one of four contrasts had contributions of more than .01 to R-square. These results could be consistent with the hypothesis of high school grade inflation for females, although a variety of other factors considered below could also produce similar findings.

More equal male–female distributions across college majors can be ruled out as an explanation for the somewhat smaller, and sometimes reversed, gender differences in the Asian American group. Contrary to the author's expectations, the distributions of quantitative versus nonquantitative majors revealed the same pattern found in other groups. With the exception of the private institution in California, substantially greater numbers of males than of females chose quantitative majors at the majority of institutions. The percentages of males versus females in the Asian American group choosing majors in the physical sciences and engineering were 47 percent versus 16 percent at the Texas university, 44 percent versus 19 percent at the university in Massachusetts, 47 percent versus 29 percent at the public institution in California, and 29 percent versus 25 percent at the private institution in California. Hence the reduced gender differences in the Asian American group as compared with the non-Latino white group before controlling for grading leniency cannot be attributed to the distribution of majors.

There is also no apparent relationship between the disparities in numbers of quantitative majors between the two sexes and the size of the underprediction of females' grades among Asian American students before controlling for grading leniency. For example, the proportion of male and female students in quantitative majors was nearly equal at the private institution in California (29 percent among males and 25 percent among females) and quite discrepant at the Texas university (47 percent among males versus 16 percent among females). Yet freshman grades (uncorrected for grading leniency) of females were underestimated by nearly the same amount (0.051 and 0.056 grade-point units at the Texas and private California institutions, respectively) when the males' equation for the SAT-M-only model was used. The differences between predicted and actual values for the standard model (leaving out MAJSCAL) were essentially zero at both institutions (−0.021 and 0.019 grade-point units at the Texas and private California universities, respectively).

Two plausible hypotheses to explain the smaller gender differences in the Asian American group come to mind. One is that male Asian American students may have such high levels of motivation and conscientiousness that they match females in their study habits, unlike males in other groups. A second hypothesis is that female Asian American students who are bilingual may have a lesser advantage in essay writing than non-Latino white female students. These hypotheses cannot be verified in the present study because there is no information available on students' motivation and study habits or on essay writing skills. However, they may be worthy of future exploration.

Sources of Gender Differences to Explain Variations Across Universities

In an earlier analysis of these data using dummy-coding of college majors that were defined in the same way for all institutions (Pennock-Román 1990), a significant intercept difference was found for the non-Latino white group at the Texas university (standard model). In the present study, controls for college major specifically tailored to each institution were used to see if better controls would eliminate this effect. Although the addition of the variable MAJSCAL did reduce gender differences, they remained statistically significant at the Texas institution. Moreover, the amount of underprediction of females' grades using the males' parameter estimates for the standard model including MAJSCAL (0.146 grade-point units) remained fairly large compared with that found in other studies of this type.

Any number of factors not controlled for in the present investigation could account for the remaining differences in prediction for males and females at the Texas institution, but there is insufficient information available about courses and the social climate at this university to evaluate which factors are more involved. Perhaps there are more essay-type examinations in the freshman year at the Texas university that give females an advantage. It is possible that there is more flexibility in nonmajor courses than at other universities, and that the Texas university females take more leniently graded courses outside their majors than males do. Maybe the university environment provides more encouragement for academic pursuits by females than by males, who might be distracted by nonacademic pursuits (e.g., football). Although females have been found to be more conscientious in their study habits and more willing to seek help with their studies than are males in general, it is not clear why these factors would be more salient at the Texas institution than at the other three universities. All four institutions in this study can be expected to have fairly demanding curricula because they are selective, major research universities. However, the Texas institution has the largest student body (the large *N* for the university in Massachusetts reflects the inclusion of two freshman classes). This suggests a hypothesis that could be tested in future research. Perhaps gender differences in grades are larger at universities with impersonal, demanding environments because females have better coping skills.

Conclusions

In the majority of studies of gender differences in the prediction of college grades, the analyses focus on the joint prediction achieved by the combination of high school grades and SAT scores. In the present research, like that of Ramist et al. (1994) and a few others, differential prediction was examined for each individual predictor and the usual combination with the objective of exploring sources of possible gender differences. The results showed little evidence that female students' high school grades were inflated in non-Latino white, African American, and Latino American groups; some results that can be interpreted as weak evidence for inflated high school grades were found for the Asian American group, however. On the other hand, the model using SAT-M as the only predictor consistently underestimated, on average, the college grades of females in all groups, even after controlling for college major. Although there was underprediction on average, much variation among individual female students occurred and the grades of some were actually overpredicted. The consistent differences between the actual average grades of females and those predicted by the males' equations were not always associated with statistically significant differences in intercepts and slopes. Nevertheless, in several instances gender effects in the SAT-M-only model, even after controlling for grading leniency, were appreciably large, accounting for more than one percent of the variance. The combination of the standard predictors had considerably higher accuracy of prediction than any individual variable considered separately, and the amount of underprediction of females' grades was usually small and not statistically significant. Thus, the standard model was the best choice, as found in many previous investigations.

Although the labor-intensive methods employed by Ramist, Lewis, and McCamley (1994), Elliot and Strenta (1988), Stricker, Rock, and Burton (1991), and Young (1990, 1991), which depend on individual course grades, are more precise for adjusting grades for leniency and improving the reliability of predicted course grades, the procedure proposed here appears to be a much easier, more practical method to control for variations in grading standards by fields of study because it avoids the analysis of transcript data and the problems of groups of unequal size across courses. This procedure improved the accuracy of prediction of freshman grades, but it is not possible to know in this study how its effectiveness compares with the control for grading leniency achieved with the other methods. If the same data were analyzed with several methods, a comparative analysis of the effectiveness of each could be done. A comparative analysis of self-reported majors from the SDQ versus institutional data on majors would also be useful.

What is known is that the improvement in prediction was seen not only in the original groups on which the classification of majors was made, but it was also cross validated in other groups at the same universities that were not involved in any way in the categorization of majors. The correction for grading leniency reduced, but did not completely eliminate, the underprediction of female students' grades by the males' equations among non-Latino white students, particularly at the Texas university. The reduction in the amount of underprediction when MAJSCAL was added was smaller in the other groups. This small residual underprediction was consistent with past research controlling for grading leniency.

Some trends in grading leniency associated with fields of study were consistent across universities (e.g., engineering majors were graded by tougher standards) and confirmed findings from past studies. Nevertheless, there were considerable variations across universities in some fields of study. The categorization of majors used here apparently had no greater advantage in reducing gender differences as compared with the dummy-variable approach to the coding of majors into four broad categories used in the earlier analyses (Pennock-Román 1990). The evidence from several studies (Elliott and Strenta 1988; Goldman and Hewitt

1975; Goldman et al. 1974; Pennock-Román 1990) suggests that the most important distinction to make is that between quantitative and nonquantitative majors. A third category should perhaps be created to distinguish the biological sciences from the quantitative sciences and other fields because grading leniency effects in biological science fields were less consistent.

Comparing actual grades and those predicted from the males' equations, the African American group showed the most underprediction of female students' grades whereas the Asian American group showed the least underprediction. These analyses suggest that combining all racial/ethnic groups into a single group for the study of gender differences may reduce the comparability of results across universities because the composition of the single group may vary greatly across institutions. Contrary to the author's expectation, male and female Asian American students were not more equally distributed among quantitative majors as compared with other groups. It is proposed that future studies explore whether or not Asian American male and female students are more equally matched in study habits and essay writing skills than males and females in other racial/ethnic categories.

References

Breland, H. M., and P. A. Griswold. 1982. "Use of a Performance Test as a Criterion in a Differential Validity Study." *Journal of Educational Psychology* 74: 713–21.

Bridgeman, B. 1989. *Comparative Validity of Multiple-Choice and Free-Response Items on the Advanced Placement Examination in Biology* (College Board Report No. 89-2). New York: College Entrance Examination Board.

Bridgeman, B., and C. Wendler. 1989. *Prediction of Grades in College Mathematics Courses as a Component of the Placement Validity of SAT-Mathematics Scores* (College Board Report 88-9). New York: College Entrance Examination Board.

Bridgeman, B., and C. Lewis. 1991. *Sex Differences in the Relationship of Advanced Placement Essay and Multiple-Choice Scores to Grades in College Courses* (ETS Research Report RR-91-48). Princeton, N.J.: Educational Testing Service.

Bridgeman, B., and C. Wendler. 1991. "Gender Difference in Predictors of College Mathematics Performance and in College Mathematics Course Grades." *Journal of Educational Psychology* 83: 273–84.

Clark, M. J., and J. Grandy. 1984. *Sex Differences in the Academic Performance of Scholastic Aptitude Test Takers* (College Board Report 84-8; ETS Research Report 84-43). New York: College Entrance Examination Board.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences.* (2nd ed.). Hillsdale, N.J.: Erlbaum.

College Board. 1988. *1988 College Bound Seniors: National Ethnic/Sex Data.* New York: College Entrance Examination Board.

Ekstrom, R. B., M. E. Goertz, and D. A. Rock. 1988. *Education and American Youth: The Impact of the High School Experience.* New York: Falmer.

Elliott, R., and A. C. Strenta. 1988. "Effects of Improving the Reliability of the GPA on Prediction Generally and on Comparative Predictions for Gender and Race Particularly." *Journal of Educational Measurement* 25: 333–47.

Gamache, L. M., and M. R. Novick. 1985. "Choice of Variables and Gender Differentiated Prediction Within Selected Academic Programs." *Journal of Educational Measurement* 22: 53–70.

Goldman, R. D., and E. N. Hewitt. 1975. "Adaptation-Level as an Explanation for Different Standards in College Grading." *Journal of Educational Measurement* 12: 149–61.

Goldman, R. D., and B. N. Hewitt. 1976. "Predicting the Success of Black, Chicano, Oriental, and White College Students." *Journal of Educational Measurement* 13: 109–17.

Goldman, R. D., D. E. Schmidt, B. N. Hewitt, and R. Fisher. 1974. "Grading Practices in Different Grading Fields." *American Educational Research Journal* 24: 287–310.

Goldman, R. D., and M. H. Widawski. 1976. "A Within-Subjects Technique for Comparing College Grading Standards: Implications in the Validity of the Evaluation of College Achievement." *Educational and Psychological Measurement* 36: 381–90.

Grandy, J. 1987a. *Ten-Year Trends in SAT Scores and Other Characteristics of High School Seniors Taking the SAT and Planning to Study Mathematics, Science, or Engineering* (ETS Research Report RR-87-49). Princeton, N.J.: Educational Testing Service.

Grandy, J. 1987b. *Trends in the Selection of Science, Mathematics, or Engineering as Major Fields of Study among Top-Scoring SAT Takers* (ETS Research Report RR-87-39). Princeton, N.J.: Educational Testing Service.

Linn, R. L. 1982. "Ability Testing: Individual Differences, Prediction, and Differential Prediction." In *Ability Testing: Uses, Consequences, and Controversies*, ed. A. K. Wigdor and W. R. Garner, pp. 335–88. Washington D.C.: National Academy Press.

Mazzeo, J., A. Schmitt, and C. Bleistein 1993. *Sex-Related Performance Differences on Constructed-Response and Multiple-Choice Sections of the Advanced Placement Examinations* (College Board Report 92-7). New York: College Entrance Examination Board.

McCormack, R. L., and M. M. McLeod. 1988. "Gender Bias in the Prediction of College Course Performance." *Journal of Educational Measurement* 25: 321–33.

Morgan, R. 1990. *Predictive Validity Within Categorizations of College Students: 1978, 1981, and 1985* (ETS Research Report 90-141). Princeton, N.J.: Educational Testing Service.

Mullis, I. V. S., and L. Jenkins. 1988. *The Science Report Card: Elements of Risk and Recovery* (Report No. 17-S-01). Princeton, N.J.: Educational Testing Service.

National Research Council. 1987. *Summary Report 1986: Doctorate Recipients from United States Universities.* Washington, D.C.: National Academy Press.

Pennock-Román, M. 1990. *Test Validity and Language Background: A Study of Hispanic American Students at Six Universities.* New York: College Entrance Examination Board.

Ramist, L., C. Lewis, and L. McCamley. 1994. *Student Group Differences in Predicting College Grades* (College Board Report 93-1; ETS Research Report 94-27). New York: College Entrance Examination Board.

Sawyer, R. 1986. "Using Demographic Subgroup and Dummy Variable Equations to Predict College Freshman Grade Average." *Journal of Educational Measurement* 23: 131–45.

Strenta, A. C., and R. Elliott. 1987. "Differential Grading Standards Revisited." *Journal of Educational Measurement* 24: 281–91.

Stricker, L. J., D. A. Rock, and N. W. Burton. 1991. *Sex Differences in SAT Predictions of College Grades* (College Board Report 91-2; ETS Research Report 91-38). New York: College Entrance Examination Board.

Wilder, G. Z., and K. Powell. 1989. *Sex Differences in Test Performance: A Survey of the Literature* (College Board Report No. 89-3; ETS Research Report 89-4). New York: College Entrance Examination Board.

Willingham, W. W. 1985. *Success in College: The Role of Personal Qualities and Academic Ability.* New York: College Entrance Examination Board.

Young, J. D. 1990. "Adjusting the Cumulative GPA Using Item Response Theory." *Journal of Educational Measurement* 27: 175–86.

Young, J. D. 1991. "Gender Bias in Predicting College Academic Performance: A New Approach Using Item Response Theory." *Journal of Educational Measurement* 28: 37–47.

Appendix

Additional Notes on the Derivation of MAJSCAL

As explained in the Method section, the variable MAJSCAL was created to reflect the grading leniency in courses in various categories of majors. It was based on the mean standardized residuals from the within-gender regressions of FGPA on HSGPA, SAT-V, and SAT-M for students in each category. Standardized residuals (residuals divided by their respective, within-gender standard errors) were used rather than the raw residuals for two reasons.

First, standardized residuals compensated for possible differences between males and females in the variance of residuals; gender differences in the variance of residuals would have affected the interpretation of the size of a mean residual. Consistently, past studies have found smaller residual variances for females because preadmission measures tend to be more highly correlated with FGPA for females (Linn 1982; Morgan 1990; Ramist et al. 1994; Sawyer 1986). If no correction were made for possible differences in residual variance by gender, the residuals for males would tend to have more extreme values than those for females. Thus, the mean residuals for major categories dominated by males would be larger in absolute value and the mean residuals for major categories dominated by females would be smaller in absolute value. In a sense, these means would not be on the same scale and they would still reflect gender effects (which we are trying to separate as much as possible from grading leniency).

The second reason for standardizing the residuals was to give their distance from zero more interpretable units. One cannot know what is a relatively large or a relatively small raw residual (in absolute value) without examining the entire distribution of raw residuals for that regression analysis.

For practical reasons, mean residuals for each major were grouped into intervals and then whole negative or whole positive values were assigned to MAJSCAL for students in that category of major according to the interval for that student's major (see Method). The use of whole numbers facilitated the key entry of values of MAJSCAL to the data set each time the categorization was revised. As explained in the Method section, the grouping for infrequent categories of majors was an iterative process involving several revisions of MAJSCAL values each time the grouping of majors was changed. If MAJSCAL had been defined to be exactly equal to the mean residual, many more revisions would have had to be made at each iteration, sometimes for only a few tenths of a point change in the means. Admittedly, there would have been an advantage in having MAJSCAL values set equal to the exact value of the mean residuals, in that the control for grading leniency would have been slightly more accurate in the non-Latino white group. However, such an exact grading-leniency rating would still be only an approximation of the grading-leniency rating in other racial/ethnic groups; it is not likely that using the exact means would have added any greater precision for controlling grading leniency in any group other than the one actually used to derive MAJSCAL. In sum, the extra precision obtained for just one group by the use of exact mean residuals did not seem worth the extra effort involved in adding MAJSCAL to the data set, either by key entry or by programming the assignment of MAJSCAL values for each university.